Course: CS699 A2 – Data Mining

Assignment: Project Assignment No.1 – Proposal

Team: Mike Zhong, Jan Allemann

# Dataset

The dataset, obtained from <https://www.redfin.com>. Redfin is a commercial real estate listing company that aggregates real estate data for consumer and business use. This particular data set was obtained by performing a search with no filters on the Boston, Massachusetts area and expanding the scope until a sufficient number of listings were included. The data set contains 350 tuples with 27 attributes each. The attributes are described in the following table:

|  |  |  |
| --- | --- | --- |
| Attribute | Format | Description |
| SALE.TYPE | Categorical | They are all MLS listing |
| SOLD.DATE | Datetime | N/A |
| PROPERTY.TYPE | Categorical | Categorical nominal |
| ADDRESS | String | House number and street name |
| CITY | String | Categorical nominal |
| STATE.OR.PROVINCE | String | Categorical nominal |
| ZIP.OR.POSTAL.CODE | String | Categorical nominal |
| PRICE | Integer | Continuous float |
| BEDS | Integer | Discrete integers |
| BATHS | Integer | Discrete ordinal |
| LOCATION | String | Categorical nominal |
| SQUARE.FEET | Integer | Discrete integers |
| LOT.SIZE | Integer | Discrete integers |
| YEAR.BUILT | Integer | Discrete integers |
| DAYS.ON.MARKET | Integer | Discrete integers |
| X..SQUARE.FEET | Integer | Continuous float |
| HOA.MONTH | Integer | Discrete integers |
| STATUS | Categorical | Categorical nominal |
| NEXT.OPEN.HOUSE.START.TIME | Datetime | Mostly empty |
| NEXT.OPEN.HOUSE.END.TIME | Datetime | Mostly empty |
| URL | String | URL to home on redfin |
| SOURCE | Categorical | Categorical nominal |
| MLS | Integer | Unique ID |
| FAVORITE | Binary | User specific variable |
| INTERESTED | Binary | User specific variable |
| LATITUTDE | Numerical | Continuous float |
| LONGITUDE | Numerical | Continuous float |

# Classification Goal

The class attribute is the PRICE, which is a continuous dollar value variable. To convert this into categorical ordinal data, we will bin the list prices using an equal width binning process. The widths of the bins is yet to be determined. The application of this classification will be suggest list prices for sellers.

The goal of the project is, to predict the PRICE category.